

633-1

MEMORANDUM

RM-3591-PR

APRIL 1963

401 393

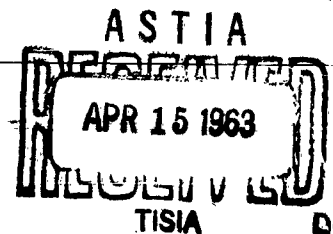
CATALOGED BY ASTIA
AS AD NO. 401393

INSTRUMENTAL VARIABLES IN
FACTOR ANALYSIS

Albert Madansky

PREPARED FOR:

UNITED STATES AIR FORCE PROJECT RAND



The RAND Corporation

SANTA MONICA • CALIFORNIA

MEMORANDUM

RM-3591-PR

APRIL 1963

**INSTRUMENTAL VARIABLES IN
FACTOR ANALYSIS**

Albert Madansky

This research is sponsored by the United States Air Force under Project RAND—contract No. AF 49(638)-700 monitored by the Directorate of Development Planning, Deputy Chief of Staff, Research and Development, Hq USAF. Views or conclusions contained in this Memorandum should not be interpreted as representing the official opinion or policy of the United States Air Force.

The **RAND** *Corporation*

1700 MAIN ST • SANTA MONICA • CALIFORNIA

PREFACE

This Memorandum points out the connection between some existing methods of factor-analysis which do not have an associated communality problem and the use of instrumental variables in econometric errors-in-variables models. It should be of particular interest to psychometricians, and of general interest to econometricians and statisticians.

SUMMARY

The factor-analysis model is rewritten as a system of linear structural relations with errors in variables. The method of instrumental variables is applied to this revised form of the model to obtain estimates of the factor loading matrix. The relation between this method and interbattery analysis, proportional profile analysis, and canonical factor-analysis is pointed out. In addition, an estimation procedure based on replicated sampling different from proportional profile analysis is given.

CONTENTS

PREFACE.	iii
SUMMARY	v
Section	
1. INTRODUCTION.	1
2. THE "ERRORS-IN-VARIABLES" MODEL.	3
3. INSTRUMENTAL VARIABLES AND INTERBATTERY ANALYSIS.	4
4. PROPORTIONAL PROFILES.	8
5. REPLICATED SAMPLING.	9
6. CANONICAL AND INSTRUMENTAL VARIABLES. . .	11
References.	15

INSTRUMENTAL VARIABLES IN FACTOR ANALYSIS

1. INTRODUCTION

The factor analysis model [2] is given by

$$(1) \quad Z = FX + U,$$

where Z is a vector of n components (n test scores, say), X is a vector of $r(\leq n)$ components (the common factor scores), U is an n -vector (the unique part of the test scores) and F is an $n \times r$ matrix (of factor loadings). It is assumed that U and X are independent random vectors, with $E U = E X = 0$, $E U U' = D$, a diagonal matrix, and $E(X X') = M$. From these assumptions, one deduces that the covariance matrix E of Z is given by

$$(2) \quad E = E Z Z' = F M F' + D.$$

After assuming that $M = I$, the identity matrix, the usual attack in determining F (or estimating F if E is unknown and only estimated) is to proceed from this equation, or equivalently from the equation corresponding to (2) based on the correlation matrix of Z . Certainly if D were known, the problem of finding F would be trivial, as it is given, up to rotation, by "factoring" the matrix $E - D$. Thus, the usual factor-analytic techniques can be characterized as attempts either at successively better approximating

D (or equivalently the diagonal terms of FF' , i.e., the communalities) or at assuming away the problem of estimating D by imposing additional restrictions on the problem which, in addition, eliminate the indeterminacy of F due to rotation (for example, the restriction that $F'D^{-1}F$ be diagonal, as in Rao's work ([16])).

In this Memorandum we shall instead proceed directly from equation (1) to the problem of estimating F. We shall see that the problem is equivalent to that of fitting a linear relation when both the independent and dependent variables are subject to error, as has been well known to some factor analysts, notably Burt [3] (see also [6]), and thus that the difficulty in estimating F without some additional restrictions on the problem (such as Rao's alluded to above) is just a reflection of the difficulties in the equivalent problem (see [14] for a catalogue of these). We shall in particular look into one of the approaches taken by econometricians when confronted with the "errors-in-variables" problem, namely the use of instrumental variables, and see that in essence this approach is what is being taken by Tucker and Gibson ([9], [10], [11], and [20]) in the interbattery method, but that they are working with equation (2), in the tradition of factor analysts, rather than with equation (1). We shall then point out connections between this approach and that of proportional profile analysis, analysis of variance in factor analysis, and canonical factor analysis.

2. THE "ERRORS-IN-VARIABLES" MODEL

As the rank of F is r , there is some $r \times r$ submatrix of F which is nonsingular. Suppose for convenience that it is the submatrix consisting of the first r rows. Write

$$Z = \begin{bmatrix} Z_{(1)} \\ Z_{(2)} \end{bmatrix} = \begin{bmatrix} F_{(1)} \\ F_{(2)} \end{bmatrix} X + \begin{bmatrix} U_{(1)} \\ U_{(2)} \end{bmatrix},$$

where $Z_{(1)}$, $F_{(1)}$, and $U_{(1)}$ have r rows. Then

$$Z_{(1)} = F_{(1)} X + U_{(1)},$$

or

$$X = F_{(1)}^{-1} (Z_{(1)} - U_{(1)}).$$

Thus

$$Z_{(2)} = F_{(2)} X + U_{(2)} = F_{(2)} F_{(1)}^{-1} (Z_{(1)} - U_{(1)}) + U_{(2)}.$$

Let $B = F_{(2)} F_{(1)}^{-1}$. Then the above equation is a linear structural relation $Z_{(2)}^* = BZ_{(1)}^*$ between the "true" variables $Z_{(1)}^* = Z_{(1)} - U_{(1)}$ and $Z_{(2)}^* = Z_{(2)} - U_{(2)}$, where the respective true variables are observed with "error" $U_{(1)}$ and $U_{(2)}$. As the matrix F is unique up to a right-multiplication by a nonsingular matrix, we might as well adopt as the canonical form for the matrix F the form

$$\tilde{F} = \begin{bmatrix} I \\ B \end{bmatrix},$$

for right-multiplication by the nonsingular matrix F_1 yields the original matrix of interest, F . Our problem is then to estimate B .

One should parenthetically note that much of the early discussion of the identifiability of the linear structural relation with errors in variables ([15], [17], [19]) was stimulated by and directed at the factor-analysis model as viewed in this light.

3. INSTRUMENTAL VARIABLES AND INTERBATTERY ANALYSIS

Suppose now that $n - r \geq r$, and, in addition to making observations on the n items comprising Z , we also observed a vector $Z_{(3)} = Z_{(3)}^* + U_{(3)}$ of $s \geq r$ additional items, where $\xi Z_{(3)} = \xi U_{(3)} = 0$, $\xi U_{(3)} U_{(3)}'$ is a diagonal matrix, $U_{(3)}$ is uncorrelated with $Z_{(3)}^*$, U , and Z^* , and $Z_{(3)}^*$ is uncorrelated with U . In effect, we assume that our vector of test items is enlarged to dimension $n + s$, and that the usual factor-analytic assumptions hold for this enlarged vector of test items. Let

$$Z_{(3)} = F_{(3)} X + U_{(3)},$$

so that

$$Z_{(3)} = F_{(3)} F_{(1)}^{-1} (Z_{(1)} - U_{(1)}) + U_{(3)},$$

and let $G = F_{(3)} F_{(1)}^{-1}$. Then

$$\mathcal{E} Z_{(2)} Z_{(3)}' = B \mathcal{E} Z_{(3)}^* Z_{(1)}'^* G',$$

$$\mathcal{E} Z_{(1)} Z_{(3)}' = \mathcal{E} Z_{(1)}^* Z_{(1)}'^* G'.$$

Thus, if the cross-covariance matrix of $Z_{(1)}$ with $Z_{(3)}$ is of maximum rank r , then

$$B = [\mathcal{E} Z_{(2)} Z_{(3)}'] [\mathcal{E} Z_{(3)} Z_{(1)}'] \{[\mathcal{E} Z_{(1)} Z_{(3)}'] [\mathcal{E} Z_{(3)} Z_{(1)}']\}^{-1}.$$

Also

$$\mathcal{E} Z_{(2)} Z_{(1)}' = B \mathcal{E} Z_{(1)}^* Z_{(1)}'^*,$$

so that if the cross-covariance matrix of $Z_{(2)}$ with $Z_{(1)}$ is of maximum rank r , then

$$G' = \{[\mathcal{E} Z_{(1)} Z_{(2)}'] [\mathcal{E} Z_{(2)} Z_{(1)}']\}^{-1} [\mathcal{E} Z_{(1)} Z_{(2)}'] [\mathcal{E} Z_{(2)} Z_{(3)}'].$$

The vector $Z_{(3)}^*$ plays the same role in this analysis as does the "instrumental variable" ([7], [18]) in the "errors-in-variables" model. Notice that the instrumental variable satisfies

$$\mathcal{E} Z_{(3)} (Z_{(2)} - \hat{B} Z_{(1)})' = 0$$

if and only if $\hat{B} = B$ under the assumptions about rank of cross-covariance matrices made earlier. This can in fact be taken as the definition of an instrumental variable, namely a vector of items which is uncorrelated with the error variable $U_{(2)} - BU_{(1)}$.

Notice that in our canonical form the test items comprising $X_{(1)}$ form a basis for our r -dimensional factor space. The matrix B is the matrix of a linear transformation mapping r -space onto $(n-r)$ -space. Given $n + s$ test items, we can, in a fashion dual to our interpretation of the notion of instrumental variable, interpret the matrix B of factor loadings as follows. It is the matrix of that linear transformation whose images $BX_{(1)}$ in $(n-r)$ -space (which we might call "pseudo-factors") are, for the purpose of telling us the correlation between factors in $(n-r)$ -space and factors in s -space, as good as the true factors in $(n-r)$ -space.

Now let us make the identification of the vector Z with battery 1, $Z_{(1)}$ with subbattery x , $Z_{(2)}$ with subbattery y , and $Z_{(3)}$ with battery 2, in Gibson's terminology [11]. We see from this that the interbattery method of factor analysis "works" because it views the more than $3r$ observed test items in the manner indicated above, as three sets of items, one a basis for the factor space and the other two as instrumental variables relative to each other. From our point of view, the matrix

$$\begin{bmatrix} I \\ B \\ G \end{bmatrix}$$

of factor loadings is as good a mathematical solution of the problem as any other. However, there is concern over obtaining orthogonal factor matrices in psychometric work. The technique of Gibson [11] as applied to our canonical form will produce orthogonal factor matrices.

The connection between this development and the work of Albert [1] should be noted. In [1] it is shown that a sufficient condition for the identifiability of the factor matrix (up to a linear transformation) is that the number of items is at least three times the rank of the factor matrix r and that the items can be partitioned into three sets such that the cross-covariance matrices between the vectors of each of the sets be of maximum rank r .

Finally, one should note that, given N independent observations $[Z_{(1)}^i, Z_{(2)}^i, Z_{(3)}^i, i = 1, \dots, N]$ on $Z_{(1)}, Z_{(2)},$ and $Z_{(3)},$ one can estimate the cross-covariance matrix between $Z_{(\alpha)}$ and $Z_{(\beta)},$ $\alpha \neq \beta, \alpha, \beta = 1, 2, 3,$ by

$$\sum_{i=1}^N \frac{Z_{(\alpha)}^i Z_{(\beta)}^{i'}}{N},$$

and so use of these estimates in place of the $E Z_{(\alpha)} Z_{(\beta)}'$ in the above equations yields a consistent estimate of the factor matrix.

4. PROPORTIONAL PROFILES

Suppose we administer our n -item test to another independent random sample of size N . Let $Z_{(3)}^i$ denote the vector of scores of the i -th individual of the new sample, $i = 1, \dots, N$. Let

$$Z = F X + U$$

$$Z_{(3)} = F_{(3)} X + U_{(3)},$$

where F and $F_{(3)}$ are $n \times r$ matrices and $F_{(3)} = F \Delta$, in which Δ is a diagonal $r \times r$ matrix. The i -th diagonal element of Δ is the differential degree of selection of the i -th factor between samples 1 and 2. This model is Cattell's proportional profile model [4], [5].

Now write the model in Z in canonical form, so that

$$Z_{(2)} = B(Z_{(1)} - U_{(1)}) + U_{(2)}$$

and

$$Z_{(3)} = \begin{bmatrix} I \\ B \end{bmatrix} \Delta (Z_{(1)} - U_{(1)}) + U_{(3)}.$$

One can thus treat the replications as a vector of instrumental variables with $s = n$ and $G = \begin{bmatrix} I \\ B \end{bmatrix} \Delta$. Using the results of Sec. 3, one can obtain estimates of B and G , and thus of Δ , by replacing B and G in the equation

$$\Delta = [I + B' B]^{-1} [I B] G$$

by their sample estimates, where I is the $r \times r$ identity matrix. This uses more of the data in estimating Δ than does the use of just the upper $r \times r$ submatrix of the estimate of G . Once one has this, one can obtain orthogonal factor matrices using Gibson's result [8].

5. REPLICATED SAMPLING

An alternative model in which to utilize the replications of the proportional profile model is to assume that $\Delta = I$ and to analyze the data in a vector analysis of variance. For this analysis, let $Z_{(1)j}$ and $Z_{(2)j}$, $j = 1, 2$, denote the replications of $Z_{(1)}$ and $Z_{(2)}$. Let

$$\bar{Z}_{(\alpha)}^i = \sum_{j=1}^2 Z_{(\alpha)j}^i / 2 \quad \alpha = 1, 2, \quad i = 1, \dots, N,$$

and

$$\bar{Z}_{(\alpha)} = \sum_{i=1}^N \bar{Z}_{(\alpha)}^i / N, \quad \alpha = 1, 2.$$

We can now look at the following vectors of appropriate deviations:

$$C_{\alpha i} = \bar{Z}_{(\alpha)}^i - \bar{Z}_{(\alpha)} \quad \alpha = 1, 2, \quad i = 1, \dots, N,$$

$$W_{\alpha i j} = Z_{(\alpha)j}^i - \bar{Z}_{(\alpha)}^i \quad i = 1, \dots, N, \quad \alpha = 1, 2, \quad j = 1, 2.$$

Then

$$\sum_{i=1}^N \sum_{j=1}^2 W_{\alpha ij} W'_{\alpha ij} / 2(N-1) = \hat{D}_{(\alpha)}$$

is an unbiased estimate of the covariance matrix of $U_{(\alpha)}$, $\alpha = 1, 2$,

$\sum_{i=1}^N C_{1i} C'_{1i} / (N-1)$ is an unbiased estimate of $D_{(1)} + 2 \sum Z^*_{(1)} Z'^*_{(1)}$,

$\sum_{i=1}^N C_{1i} C'_{2i} / (N-1) = 2 \hat{S}_{12}$ is an unbiased estimate of

$2 \sum Z^*_{(1)} Z'^*_{(2)} = 2 B \sum Z^*_{(1)} Z'^*_{(1)}$, $\sum_{i=1}^N C_{2i} C'_{2i} / (N-1)$ is an unbiased estimate of $D_{(2)} + 2 \sum Z^*_{(2)} Z'^*_{(2)} = D_{(2)} + 2 B \sum Z^*_{(1)} Z'^*_{(1)} B'$.

Thus B can be consistently estimated from these data in a variety of ways. The simplest of these is to estimate $\sum Z^*_{(1)} Z'^*_{(1)} = S_{11}$ by

$$\sum_{i=1}^N C_{1i} C'_{1i} / (N-1) - \hat{D}_{(1)} / 2 = \hat{S}_{11}$$

next $\sum Z^*_{(1)} Z'^*_{(2)} = S_{12}$ by \hat{S}_{12} , and then estimating B by

$$\hat{B} = \hat{S}_{12} \hat{S}_{11}^{-1}.$$

The fact that the proportional-profiles model, where $\Delta = I$, and so where the degrees of selection are not at all different, can be analyzed is quite contrary to the forebodings of Gibson ([8], p. 136). In fact, the development in this section can be extended to arbitrary, but known, Δ .

6. CANONICAL AND INSTRUMENTAL VARIABLES

Harris [12] and Kaiser [13] have presented the following intuitive picture of the basis for Rao's canonical factor analysis. Consider the random $(n+r)$ -dimensional vector

$$Y = \begin{pmatrix} Z \\ X \end{pmatrix}.$$

It has covariance matrix

$$\begin{bmatrix} \Sigma & FM \\ MF' & M \end{bmatrix},$$

and the squared canonical correlations between Z and X are given by the roots of the determinantal equation

$$|FMF' - \lambda \Sigma| = 0.$$

Since $\Sigma = FMF' + D$, this is equivalent to finding roots of the equation

$$|FMF' - \Theta D| = 0 = |\Sigma + D - \lambda \Sigma| = |\Sigma - \xi D|,$$

where $\Theta = \lambda / (1 - \lambda)$ and $\xi = -1 / (1 - \lambda) = -(1 + \Theta)$. Let U be the matrix of characteristic vectors associated with the roots Θ , and let Θ be the diagonal matrix of Θ 's, where

$$(D^{-1/2} F M F' D^{-1/2}) U = U \Theta .$$

Then

$$F M F' = D^{1/2} U \Theta U' D^{1/2}$$

and, up to a linear transformation,

$$F = D^{1/2} U \Theta^{1/2} .$$

If $M = I$, then this is correct up to an orthogonal transformation, and if $F' D^{-1} F$ is diagonal, then U must be orthogonal and the above factor loadings are unique.

Let us look for a moment at what would be the perfect instrumental variable. As noted in Sec. 3, an instrumental variable is one which is uncorrelated with $U_{(2)} - B U_{(1)}$. Certainly X , which is uncorrelated with the U 's, fits this description. And the Harris-Kaiser description of canonical factor analysis points up that X can be viewed initially as an additional set of variables, i.e., as an instrumental variable.

The difficulty with using $X = Z_{(3)}$ in the equations of Sec. 3 is that X is not observable. This difficulty is circumvented in canonical factor analysis by the assumption that $F' D^{-1} F$ is diagonal. What are the implications of this assumption on our canonical form of the factor analysis problem?

Since $F' = (I B')$, we see that

$$F'D^{-1}F = D_1^{-1} + B'D_2^{-1}B = \Delta,$$

say. We also have the equation

$$\Sigma D^{-1}F = F\Delta,$$

or

$$\begin{pmatrix} \Sigma_1 D_1^{-1} + \Sigma_2 D_2^{-1} B \\ \Sigma_2' D_1^{-1} + \Sigma_4 D_2^{-1} B \end{pmatrix} = \begin{pmatrix} \Delta \\ B\Delta \end{pmatrix},$$

where

$$D = \begin{pmatrix} D_1 & 0 \\ 0 & D_2 \end{pmatrix} \text{ and } \Sigma = \begin{pmatrix} \Sigma_1 & \Sigma_2 \\ \Sigma_2' & \Sigma_4 \end{pmatrix}.$$

Since $\Sigma_1 = I + D_1$, $\Sigma_4 = BB' + D_2$, and $\Sigma_2 = B'$, the equation

$$\begin{aligned} \Delta &= D_1^{-1} + B'D_2^{-1}B = \Sigma_1 D_1^{-1} + \Sigma_2 D_2^{-1}B \\ &= (I + D_1) D_1^{-1} + B'D_2^{-1}B \end{aligned}$$

holds only if $D_1 = 0$, so that D_1^{-1} has no meaning.

Also,

$$\begin{aligned} B\Delta &= BD_1^{-1} + BB'D_2^{-1}B = \Sigma_2' D_1^{-1} + \Sigma_4 D_2^{-1}B \\ &= B'D_1^{-1} + (BB' + D_2) D_2^{-1}B \end{aligned}$$

holds only if $B = 0$. Thus when the assumptions of canonical factor analysis hold, we see this immediately in our canonical form, as then the first r items are the factors and are observed without error.

REFERENCES

1. Albert, A. A., "The Minimum Rank of a Correlation Matrix," Proc. Nat. Acad. Sci., 30 (1944), pp. 144-146.
2. Anderson, T. W., and Herman Rubin, "Statistical Inference in Factor Analysis," Proc. Third Berkeley Symposium on Math. Stat. and Prob., Vol. V, J. Neyman, ed., University of California Press, 1956, pp. 111-150.
3. Burt, C., "Alternative Methods of Factor Analysis and their Relations to Pearson's Method of 'Principle Axes,'" Br. J. Psych., Stat. Sec., 2 (1949), pp. 98-121.
4. Cattell, R. B., "Parallel Proportional Profiles and other Principles for Determining the Choice of Factors by Rotation," Psychometrika, 9 (1944), pp. 267-283.
5. Cattell, R. B., and A. K. S. Cattell, "Factor Rotation for Proportional Profiles: analytical solution and an example," Br. J. of Stat. Psych. 8 (1955), pp. 83-92.
6. Creasy, M. A., "Analysis of Variance as an Alternative to Factor Analysis," J. Royal Stat. Soc., Series B, 19 (1957), pp. 318-325.
7. Geary, R. C., "Determination of Linear Relations Between Systematic Parts of Variables with Errors of Observation the Variances of which are unknown," Econometrica, 17 (1949), pp. 30-59.
8. Gibson, W. A., "Proportional Profiles and Latent Structure," Psychometrika, 21 (1956), pp. 135-144.
9. Gibson, W. A., "Remarks on Tucker's Inter-Battery Method of Factor Analysis," Psychometrika, 25 (1960), pp. 19-25.
10. Gibson, W. A., "A Retraction of Inter-Battery Factor Analysis," Psychometrika, 26 (1961), pp. 451-52.
11. Gibson, W. A., "An Asymmetric Approach to Multiple-Factor Analysis," Br. J. Stat. Psych., 14 (1961), pp. 97-107.

12. Harris, C. W., "Some Rao-Guttman Relationships," Psychometrika 27 (1962), pp. 247-64.
13. Kaiser, H. F., and John Caffrey, "A Psychometric Method of Factor Analysis," unpublished manuscript, March 1962.
14. Madansky, A., "The Fitting of Straight Lines when both Variables are Subject to Error," Jour. Am. Stat. Assoc., 54 (1959), pp. 173-205.
15. Neyman, J., "Remarks on a paper by E. C. Rhodes," J. Royal Stat. Soc., 100 (1937), pp. 50-57.
16. Rao, C. R., "Estimation and Tests of Significance in Factor Analysis," Psychometrika, 20 (1955), pp. 93-111.
17. Rhodes, E. C., "On Lines and Planes of Closest Fit," The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, Series 7, Vol. 3 (1927), pp. 357-364.
18. Sargan, J. D., "The Estimation of Economic Relationships using Instrumental Variables," Econometrica, 26 (1958), pp. 393-415.
19. Thomson, G. H., "A Hierarchy without a General Factor," Br. J. Psych., 8 (1916), pp. 271-281.
20. Tucker, L. R., "An Inter-Battery Method of Factor Analysis," Psychometrika, 23 (1958), pp. 111-136.